**Bike Sharing Demand Prediction**

**Samara Reddy**

**Data Science Trainee,**

**Alma Better, Bangalore.**

**Contents**

1. **ABSTRACT**
2. **INTRODUCTION**
3. **PROBLEM STATEMENT**
4. **METHODOLOGY**

**4.1) Dataset Information**

**4.2) Attribute Information**

**4.3) Data Pre-processing**

**4.4) Exploratory Data Analysis**

**4.6) Hyperparameter Tuning**

**4.7) Machine Learning Models**

**4.7.1) Linear Regression**

**4.7.2) Lasso Regression**

**4.7.3) Ridge Regression**

**4.7.4) Random Forest Regressor**

1. **EVALUATING ML MODELS**

**RESULTS**

**ABSTRACT**

A Bike-sharing system is a shared transport service in which bikes are made available for shared use to individuals on a short-term basis for a price. Many bike share systems allow people to borrow a bike from a "dock" and return it at another dock belonging to the same system. Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes. The first segment of this work is dedicated towards the exploratory data analysis i.e. understanding the pattern lying beneath and second part is dedicated towards applying different models and selecting the appropriate one.

**INTRODUCTION**

Bike-renting systems allow anyone to hire bike from one of the city’s numerous automated rental stations, ride it for a short distance, and then return it to any station in the city. Many cities across the world have recently implemented similar systems. First country to implement this model is the Portland, Oregon, following their example bike rental systems are widespread throughout the world .The ability of a rental bike system to meet the variable demand for bicycles and make it available at the time of peak of its demand. This is accomplished by a repositioning operation that involves taking bicycles from some stations and transporting them to other stations with the help of a specialized fleet of trucks. From the economical point of view, this process is highly intensive and expensive for bike sharing companies. Therefore another optimal solution is needed. To solve such a problem machine learning methods will used. To what extent we can predict bike count required for the stable supply of rental bikes using machine learning. Since the goal is to find the best model that will output highest accuracy on data set, therefore the first sub question that comes to mind is: “Which machine learning algorithm gives the highest prediction accuracy.



**PROBLEM STATEMENT**

It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

In addition, there are many features in the data such as temperature and time which can be used to predict bike counts. It would be interesting to see which feature is more importance for prediction. Therefore, a second question that comes to mind is “Which features are the most important for bike count prediction”.

**Dataset Information**

The dataset contains weather information, the number of bikes rented per hour and date information.

The time span of the dataset is 365 days from December 2017 to November 2018.

The dataset consists of 8760 rows and 14 features(columns).

Dependent Variable- Rented Bike Count and remaining were Independent variable

**Attribute Information**

* **Date** : *The date of the day, type:str*
* **Rented Bike Count** - *Number of rented bikes per hour and it is also a dependent variable, type:int*
* **Hour** - *Hour of the day ranging from 0-23, type: int*
* **Temperature (°C)**-*Temperature in Celsius, type:float*
* **Humidity(%)** - *Humidity in the air in %, type: int*
* **Wind speed (m/s)** - *Speed of the wind in m/s, type: float*
* **Visibility (10m)** - *Visibility in m, type: int*
* **Dew point temperature(°C)** - *The temperature at which the water starts to condense out of the air, type: float*
* **Solar Radiation (MJ/m2)** - Electromagnetic radiation emitted by the Sun, type: float
* **Rainfall(mm)** - Amount of rainfall in mm, type: float *italicized text*
* **Snowfall(cm)** - *Amount of snowfall in cm, type: float*
* **Seasons** - *Season of the year, type: str*
* **Holiday** - *If the day is holiday or not, type: str*
* **Functioning Day** - *Whether the day is functional or not, type:str*

**Data Pre-processing**

In this section, methods for data exploration and pre-processing will be presented so that we can make our data fit for passing into different ML models. The majority of the realworld datasets are highly susceptible to missing, inconsistent, and noisy da ta due to their heterogeneous origin. Applying algorithms on this noisy da ta would not give quality results as they would fail to identify patterns effectively. Data Processing is, therefore an important step or stage to improve the overall data quality. Duplicate or missing values may give an incorrect view of the overall statistics of data. Outliers and inconsistent data points often tend to disturb the model’s overall learning, leading to false predictions. Major Tasks in Data Pre-processing:

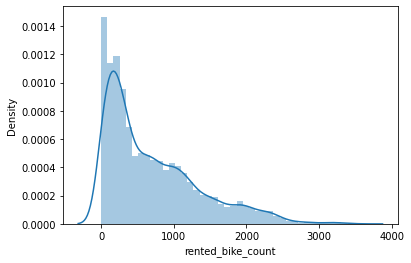
1. Data cleaning 2. Data integration 3. Data reduction 4. Data transformation

EDA or Exploratory Data Analysis is the critical process of performing the initial investigation on the data to find the anomalies in our data and shape it such that it is useful for taking some insights to sol purpose. There are certain step that we follow initially we will clean our data and make it free from such as Nan values, missing and such values that could hinder the accuacy of our analysis.

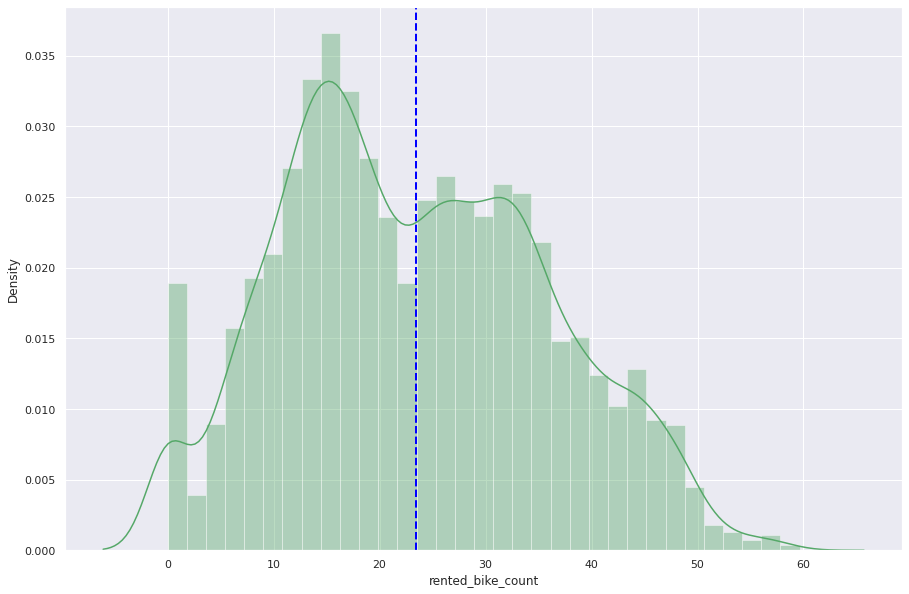
The ultimate aim of the step is to provide our ML models the best possible and clean data, for that we need to treat the outliers; duplicate values must also be taken into account. Luckily, in our dataset there were absolute zero missing values and duplicate values. Therefore data was clean in this regard.

It is crucial to handle ca during the pre-processing phase, as in the data having categorical features, namely 'hour', 'se 'holiday', 'functioning day', 'month' Machine learning can’t categorical data. Hence, converting into numerical ones. One hot encoding, creating dummy variables and many other methods can be used for categorical variables. Based on current work, dummy variables were created for each of the features.

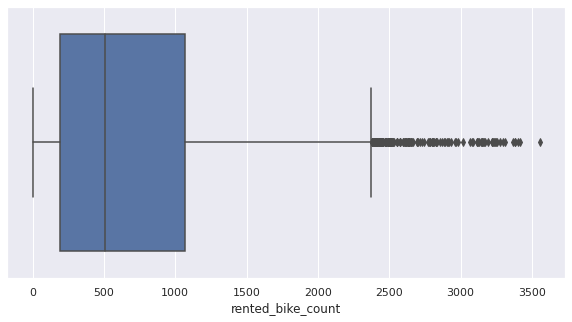
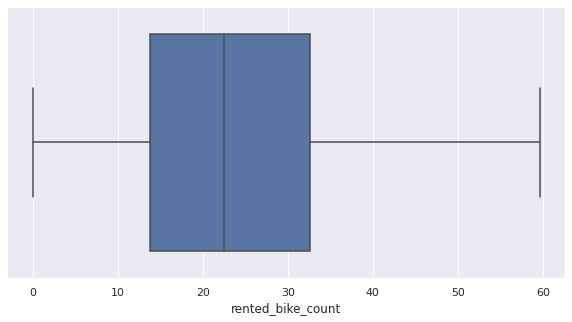
The next step is to verify that the data is normal. Target variable distribution seems to not normal distribution.



Thus, data was transformed to closely resemble a normal distribution. We have applied here square – root method to normalize the dependent variable. Initially, Log – transformed was applied but due the presence of ‘0’ at some instances. Therefore, mathematical transformer is used.



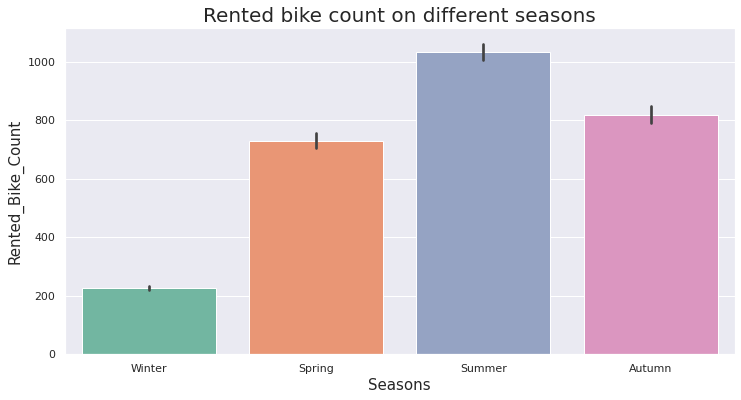
The outliers were also got treated as soon as we applied the transform.

Before After

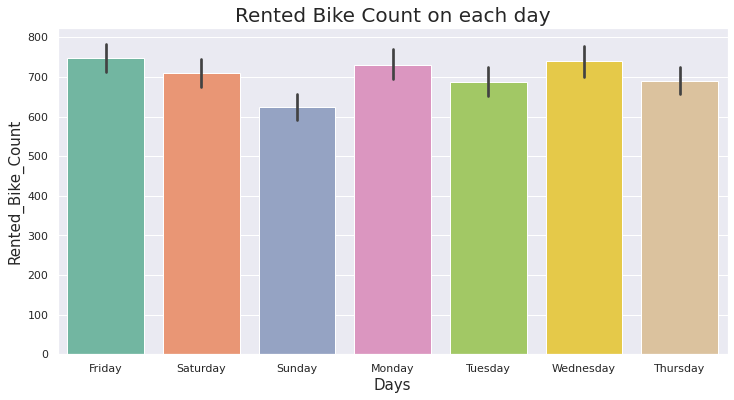
Data visualization is the most important step while doing the analysis. It is more impressive, interesting and understanding when we represent our study or analysis with the help of colours and graphics. Using visualization elements like graphs, charts, ma The outliers were also got treated as soon as we applied the transform. Data visualization is the most important step while doing the analysis. It is more impressive, interesting and understanding when we represent our study or analysis with the help of colours and graphics. Using visualization elements like graphs, charts, maps, etc., it becomes easier to understand the underlying structure, trends, patterns and relationships among variables within the dataset. The libraries use for the plotting visualisation are Seaborn. Some of the plots are show below:

**Finding relation between seasons and rented bike count**



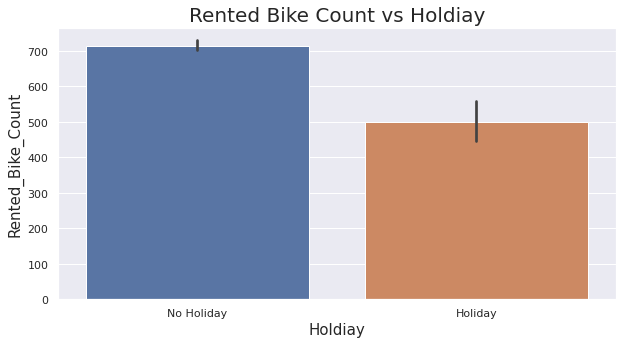
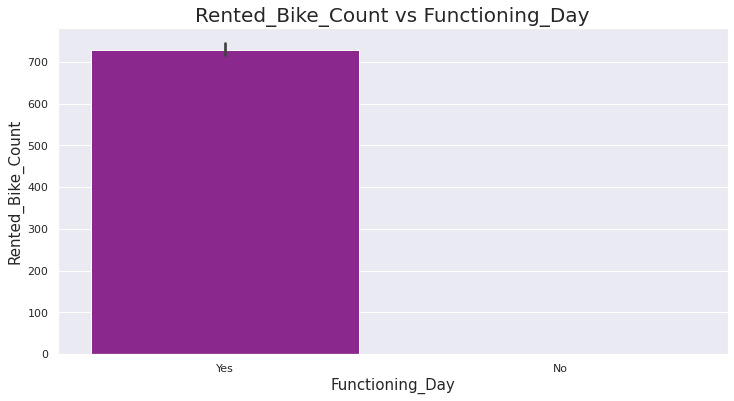
Rented bike count is significantly less in Winter whereas high on summer

**plot for the rented bike count against weekday**



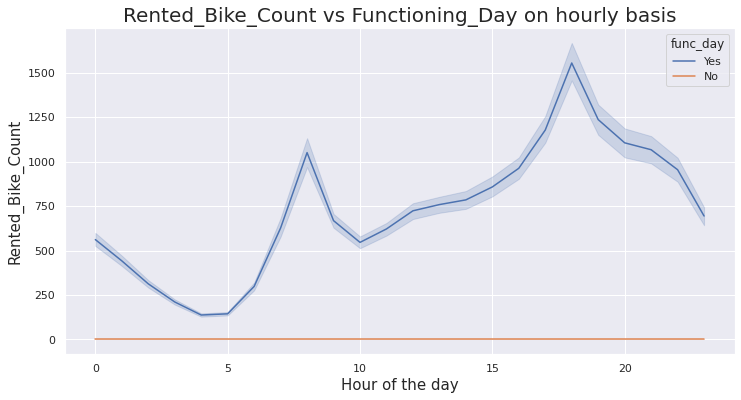
In all days, the bike count is more or less the same

Understanding the rented bike count on holiday vs non holiday and functioning day v/s a non functioning day

We see that people are renting bike more on non holidays than holidays. The rented bike count is 0 on a non functioning day.

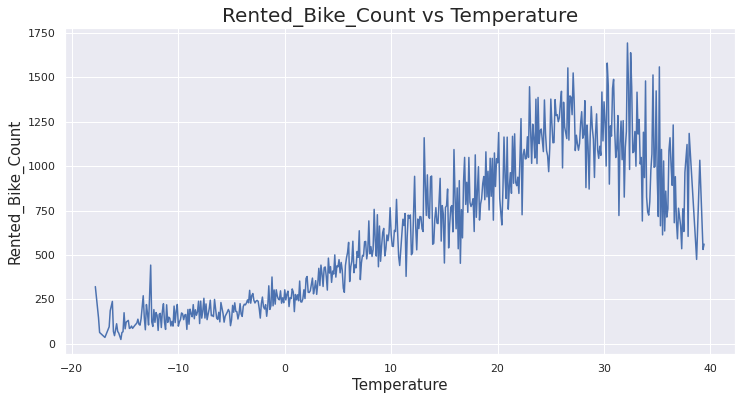
**Rented bike count on func vs non func day on an hourly basis.**



The above plot suggests that the people use bikes only on functioning day.

Also it means from company's perspective, if anyone rents bike on a day then that is a functioning day.

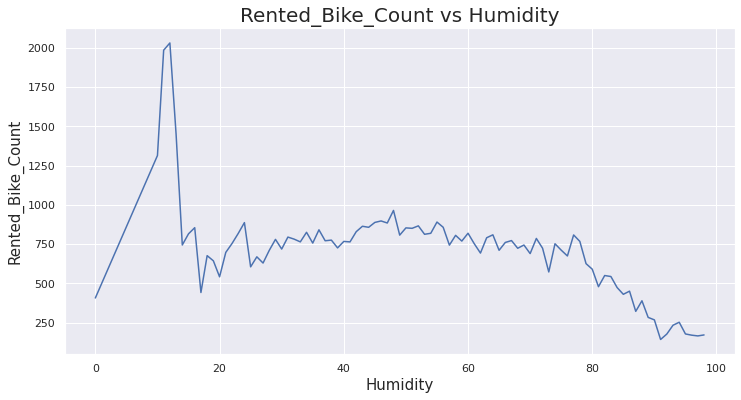
**Plotting number of bike rented vs temperature**



There is a increase in demand of rental bikes as the temperature increases.

The maximum demand is around 28-33 degree celsius.

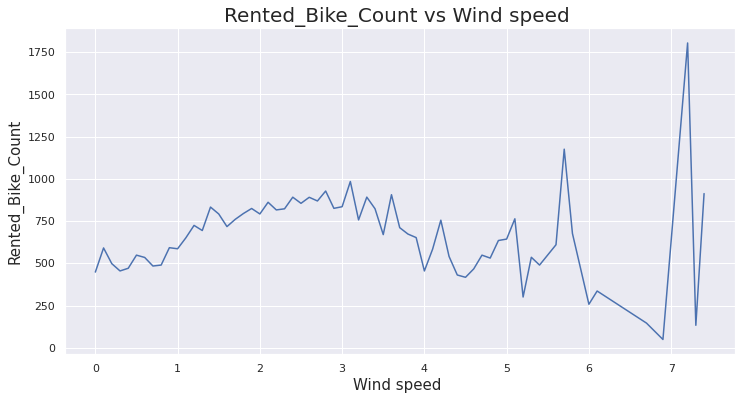
**Understanding the relation between rented bike count and humidity**



The demand for rented bikes drops as humidity increases.

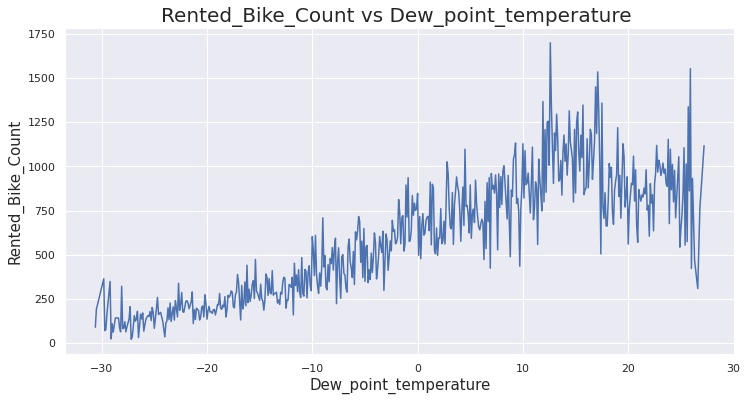
The rented bike count is maximum when humidity is around 15%.

**Understanding how wind speed affects the bike count through lineplot**



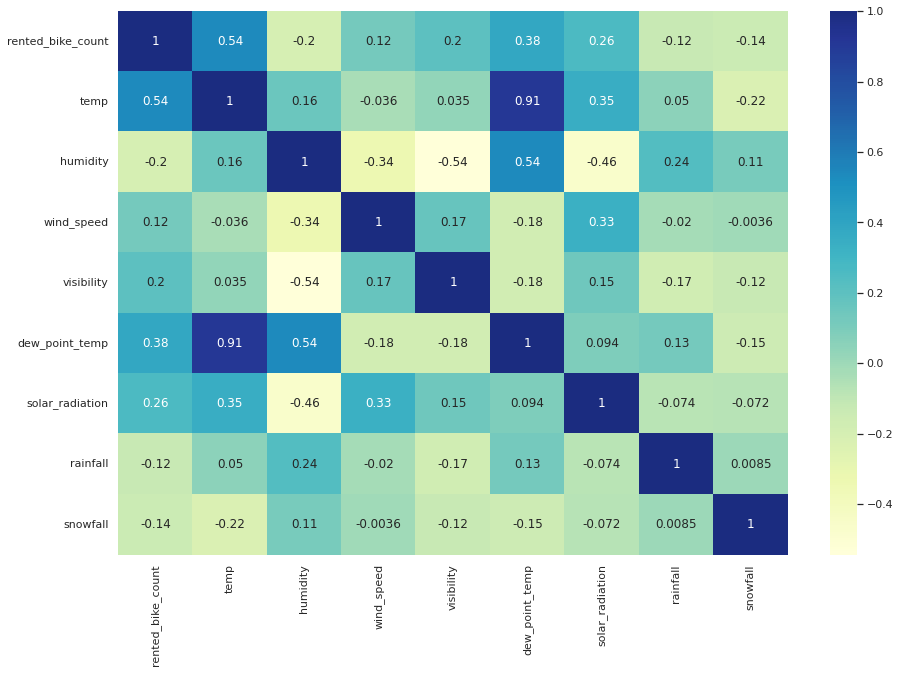
The bike count gradually decreases as the wind speed increases but its maximum at around 7m/s.

**Understanding relation between dew point temperature and rented bike count**



There is an increasing trend in rented\_bike\_count upto almost 18 degree Celsius.

**Correlation Matrix:**

****

Hour and temperature are highly correlated with bike count as compared to other features.

Rainfall and snowfall are negatively correlated with rented bike count.

Dew point temp and temperature are highly correlated with each other, we can add them together.

**Machine Learning Models:**

**Linear Regression:** Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Below shown are the Evaluation Metrics results, it represents the goodness of fit of a model.

**Test data evaluation metrics**

**MSE : 33.480918149917905**

**RMSE : 5.786269795811279**

**MAE : 4.392178273516105**

**R2 : 0.7874041916419688**

**Adjusted R2 : 0.7808974335285976**

Since, R2 we got is we can say that our model gave us good results.

**Lasso Regression:** Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso Regression uses L1 regularization technique. It is used when we have more features because it automatically performs feature selection. The word “LASSO” stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. It is a statistical formula for the regularisation of data models and feature selection.

Below shown are the Evaluation Metrics results that we got, it represents the goodness of fit of a model

**MSE : 33.48120283557305**

**RMSE : 5.786294395861055**

**MAE : 4.392172437320416**

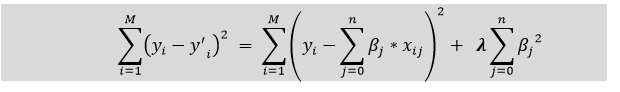
**R2 : 0.7874023839562682**

**Adjusted R2 : 0.7808955705164364**

**{'alpha': 0.0001} is best Hyperparameters that I got by GridSearchCV method.**

**Ridge Regression:** Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions. Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization.** In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called **Ridge Regression penalty**. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.

The equation for the cost function in ridge regression will be



Below shown are the Evaluation Metrics results that we got, it represents the goodness of fit of a model.

**MSE : 33.49659664299544**

**RMSE : 5.787624438661811**

**MAE : 4.39453827031332**

**R2 : 0.7873046369674306**

**Adjusted R2 : 0.7807948318599005**

**{'alpha': 1} is best Hyperparameters that I got by GridSearchCV method.**

**Random Forest Regressor:** Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.   
Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

Below shown are the Evaluation Metrics results that we got, it represents the goodness of fit of a model.

**MSE : 37.34634463300656**

**RMSE : 6.111165570740704**

**MAE : 4.4076482411074**

**R2 : 0.762859659615066**

**Adjusted R2 : 0.7556016856892175**

**{'max\_depth': 8, 'min\_samples\_leaf': 40, 'min\_samples\_split': 50, 'n\_estimators': 150}**

**Are the best Hyperparameter that I got by GridSearchCV method**

**RESULTS**

* There are not any duplicate rows and missing values in the dataset
* The rented bike count data is positively skewed and After square root transformation outliers have been removed.
* Rented bike count is significantly less in Winter whereas high on summer
* In all days, the bike count is more or less the same
* People are renting bike more on non holidays than holidays and rented bike count is 0 on a non functioning day
* People use bikes only on functioning day. Also it means from company's perspective, if anyone rents bike on a day then that is a functioning day.
* There is a increase in demand of rental bikes as the temperature increases. The maximum demand is around 28-33 degree Celsius
* The demand for rented bikes drops as humidity increases. The rented bike count is maximum when humidity is around 15%.
* The bike count gradually decreases as the wind speed increases but its maximum at around 7m/s
* Hour and temperature are highly correlated with bike count as compared to other features
* Dew point temp and temperature are highly correlated with each other, I added them together.
* Linear Regression Algorithm has got Lowest RMSE Score 5.78 and Highest Adjusted R2 score 0.78 among all the models